WIDE AREA LOAD BALANCING OF WEB TRAFFIC

ABSTRACT OF THE DISCLOSURE

Methods and apparatus are described for intelligently for intelligently routing a request to a device (e.g., replica or server). A packet is received (e.g., by the client's gateway router) from a client, and the packet has a destination identifier associated with a server. When the packet is a start packet, a tag is added to the start packet to indicate that the start packet should be forwarded to any replica that duplicates the data content of the server. The destination identifier of the start packet is stored for later use. After storing the destination identifier of the start packet and tagging the start packet, the start packet is sent to the server. When the start packet has a tag indicating that the start packet should be forwarded to any replica that duplicates the data content of the server, the start packet is encapsulated and sent to each replica associated with the server. A replica device then receives a start packet sent from the client to the server. The start packet is encapsulated. The encapsulated start packet is cracked to determine the client's address. When the replica device is active and not busy, an immediate acknowledgement packet is sent to the client in response to the received start packet. When a first acknowledgement packet associated with the start packet is received (e.g., at the client gateway router), a source identifier of the first acknowledgement packet is stored and associated with the stored destination identifier of the start packet. After storing and associating the source identifier of the first acknowledgement packet, the first acknowledgement packet is sent to the client. Subsequent packets (*e.g.*, after the start and acknowledgement packets) are then sent between the client and sender of the first acknowledgement packet.